An Interpretable Deep Learning Approach to Understand Health Misinformation Transmission on YouTube

Jiaheng Xie
University of Delaware
jxie@udel.edu

Yidong Chai
Hefei University of Technology
chaiyd@hfut.edu.cn

Xiao Liu
Arizona State University
xiao.liu.10@asu.edu

Abstract

Health misinformation on social media devastates physical and mental health, invalidates health gains, and potentially costs lives. Deep learning methods have been deployed to predict the spread of misinformation, but they lack the interpretability due to their blackbox nature. To remedy this gap, this study proposes a novel interpretable deep learning, Generative Adversarial Network based Piecewise Wide and Attention Deep Learning (GAN-PiWAD), to predict health misinformation transmission in social media. GAN-PiWAD captures the interactions among multi-modal data, offers unbiased estimation of the total effect of each feature, and models the dynamic total effect of each feature. Interpretation of GAN-PiWAD indicates video description, negative video content, and channel credibility are key features that drive viral transmission of misinformation. This study contributes to IS with a novel interpretable deep learning that is generalizable to understand human decisions. We provide direct implications to design interventions to identify misinformation, control transmissions, and manage infodemics.

1. Introduction

The misinformation transmitted on social media is detrimental to individual's physical and mental health, elevates stigmatization and hate speech, threatens precious health gains, leads to poor observation of public health measures, and even costs lives (WHO 2020). A recent study shows that more than 25% of the most viewed COVID-19 videos contain misinformation (Li et al. 2020). Among those misinformation videos, one myth – that highly concentrated alcohol consumption could disinfect the coronavirus – infiltrated the public's belief and claimed over 800 lives (Islam et al. 2020).

This study aims to predict health misinformation transmission on social media as well as to identify the driving factors of its transmission. We define the transmission of misinformation as the daily viewership of this misinformation. We leverage the social exchange theory, which is commonly used to explain human information sharing behavior, to build the theoretical foundation of this study (Liang et al. 2008). Among the social media platforms that transmit misinformation, YouTube receives the most concerning attention because of its easy-to-implement audio messages, visual presentations to spread misinformation, and extensive user base. We focus on the misinformation transmission on YouTube and attempt to predict the viewership of health misinformation videos and unveil the mechanism of their transmission.

How does a video become viral? This is one of the well-known open research questions in social media analytics. Prior work has studied the relationship between content popularity and various factors, including network actor properties, content features, and effects of complex contagion, among others (Cheng et al. 2014, Romero et al. 2011). These video analytics studies rely on deep learning methods to predict an outcome of a video, for instance, the transmission of a video in this study. Despite the premier predictive power of deep learning, its blackbox nature falls short in interpreting the driving factors, thus failing to provide proactive and implementable plans to manage the transmission of those videos. In order to manage infodemics, it is essential to not only predict misinformation transmission, but also simultaneously understand the underlying factors that drive such a transmission. This objective necessitates fine-grained interpretable deep learning methods.

An emerging stream of research that reconciles interpretability and accuracy is wide and deep learning (Cheng et al. 2016). This method incorporates an interpretable linear model in the deep learning model. The linear part and deep part are trained jointly. The linear part is capable of interpreting how the input contributes to the prediction. We define the effect interpreted by the linear part (weights of the linear part) as the main effect. Since the introduction of the wide and deep learning, numerous of its variants emerged to improve the prediction performance and minimize the biases of interpretation. Despite those efforts, the wide and deep learning and its variants fall short in producing

URI: https://hdl.handle.net/10125/79515
978-0-9981331-5-7

unbiased interpretations for deep learning predictions. When training the linear part and the deep part jointly, the deep part influences the weights of the linear part. Therefore, the final weights of the linear part (main effect) are not the total effect of the input on the output. Using those weights to interpret the prediction – as the existing interpretable methods do – introduces biases. In addition, the existing wide and deep learning and its variants neglect the interactions between the inputs, which is critical in this study. Media richness theory suggests that the interactions between the underlying factors are essential elements for explanatory information (Wheeler and Arunachalam 2009). Furthermore, the total effect of a feature on the outcome is not constant. It is dynamic when the value of the feature varies. For instance, the total effect of video duration on viewership is stronger if the video is short. This total effect gradually diminishes if the video is long. This is because when a video is very long, viewers do not have the incentive to watch it no matter its duration increases or decreases marginally or not. To address the limitations of the existing interpretable methods, we aim to propose a novel model-based interpretable method that captures the interaction effects, produces unbiased interpretation, and models the dynamic total effect of each feature. Our proposed method is called Generative Adversarial Network based Piecewise Wide and Attention Deep Learning (GAN-PiWAD).

This study makes the following contributions to data analytics methodology, information systems (IS) literature, and social media analytics. First, we develop a novel GAN-PiWAD method that is capable of simultaneously predicting and interpreting health misinformation transmission. Our method addresses the deficiencies of post-hoc interpretable methods and innovatively designs a Wasserstein generative adversarial network with gradient penalty (WGAN-GP) and an attention-based second-order component. Empirical evaluations indicate that GAN-PiWAD outperforms all the strong baseline models. Each component of GAN-PiWAD significantly contributes to the performance gain, suggesting successful design choices. The superior performance of GAN-PiWAD not only contributes to the data analytics methodology, but also offers indispensable design principles for the design science paradigm of information systems research. Our method development and evaluations prove that designing a second-order component to capture the interactions among the multi-modal inputs could boost the predictive performance of interpretable methods. Our newly-added deep generative component in our method also offers a generalizable approach to estimate the unbiased total effect of prediction tasks. Our method outlines an innovative approach to model the dynamic total effect of each factor. The complex total effect is taken into consideration when the factor value changes. The proposed GAN-PiWAD is a generalizable interpretable deep learning model that can transit to many other predictive tasks, such as user engagement prediction, product sales prediction, and project investment prediction, among others.

Second, we contribute to the computational information systems research (Abbasi et al. 2010, 2012, Fang et al. 2013, Mai et al. 2018, Saboo 2016, Stieglitz and Dang-Xuan 2013). We identified an societal problem and designed an analytics tool to predict and understand this problem. Empirical evaluations proved the efficacy of our method. The interpretation of the predictive method sets an examplar for other design science studies to not only predict an outcome, but offer invaluable interpretability as well.

Third, our findings provide social media platforms with practical implications regarding infodemics control. Our method is capable of identifying misinformation sources that are on the edge of abruption. To help these platforms design actionable intervention plans, our model interpreted the prediction results. The interpretation indicates that video description features, negative video content, and channel credibility are critical features that drive the transmission of misinformation. Social media platforms could leverage our method to actively monitor these factors and prevent the transmission of misinformation.

2. Literature Review

2.1. Theoretical Background

In order to understand the information transmission behavior, extensive studies in IS leveraged the social exchange theory. Social exchange theory is widely applied to explain individual behaviors across various domains, including information sharing (Kankanhalli et al. 2005), information technology adoption (Gefen and Keil 1998), consumer behavior (Ba and Pavlou 2002, Shiau and Luo 2012), and behavior in online communities (Jin et al. 2010). Social exchange theory is a broad conceptual paradigm that spans a number of social scientific disciplines, such as management, social psychology, and organization science (Cropanzano et al. 2017). Despite its name, it is not a single theory but a family of conceptual models. All social exchange theories treat social life as a series of sequential transactions between groups as well as between individuals. Resources are exchanged through a process of reciprocity, whereby one party tends to repay the good (or sometimes bad) deeds of another party (Gergen 1969).

In this research, we leverage the social exchange theory as our theoretical foundation to guide the feature selection of our interpretable computational model. Based on prior literature on the social exchange theory, we identify four essential constructs of a social exchange in the context of health misinformation transmission on social media platforms: cost, rewards, status, and emotion.

2.2. Video Analytics Features

In this study, we examine the factors from prior literature according to a comprehensive video features from YouTube and investigate the respective as well as joint effects of these factors. Figure 1 depicts the theoretical framework we propose in the YouTube health misinformation research context. This work also adds to the existing literature in public health and health communication research by developing a research framework to explain the transmission of health misinformation and providing comprehensive guidelines for platforms and users to mitigate the harm of misinformation.

In order to address these limitations, the interpretable component needs to be cohesively embedded in the prediction model, which is also called model-based interpretable methods. The model-based interpretable methods have a self-contained structure that not only makes accurate predictions with a single objective function, but also precisely characterizes the relationship between the input features and the outcome. State-of-the-art deep learning models make predictions through a deep neural network. The nonlinear relations between the input and output are captured in the hidden layers of this network. Because of the depth of such architecture, the relations in the hidden layers are not interpretable. To overcome this limitation, Cheng et al. (2016) proposed the wide and deep learning (W&D) that trains an interpretable linear component jointly with a deep neural network. The wide component of this method is a linear model. This wide component produces a weight for each feature (main effect) to interpret the prediction. The second joint component is a deep neural network. This deep component models high-order relations in the hierarchical network to improve prediction accuracy. The wide and deep learning combines the wide and deep components leveraging the weighted sum.

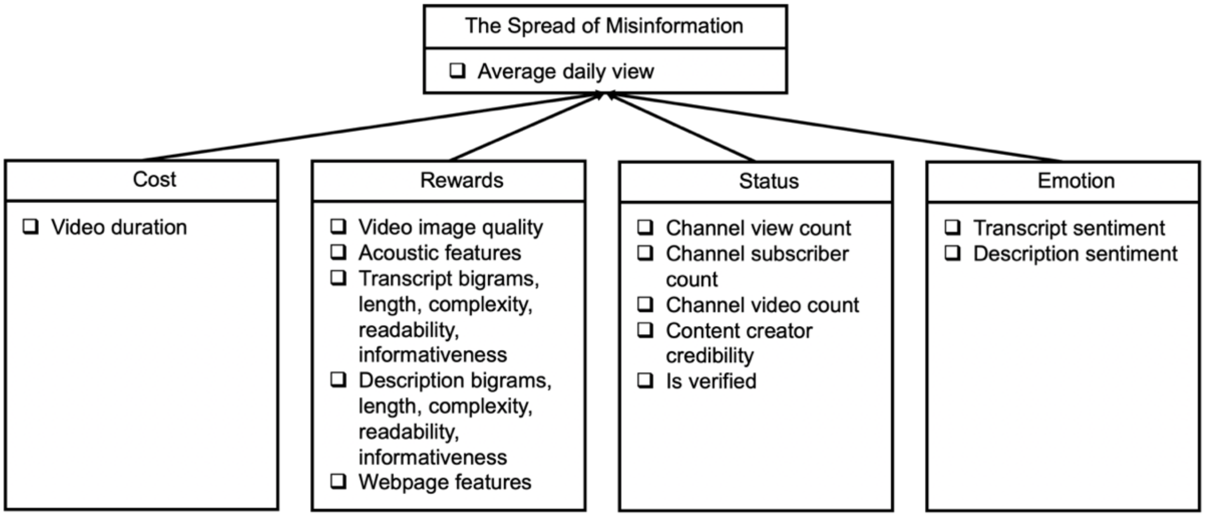
The Spread of Misinformation
❑ Average daily view
Cost
❑ Video duration
Rewards
❑ Video image quality
❑ Acoustic features
❑ Transcript bigrams, length, complexity, readability, informativeness
❑ Description bigrams, length, complexity, readability, informativeness
❑ Webpage features
Status
❑ Channel view count
❑ Channel subscriber count
❑ Channel video count
❑ Content creator credibility
❑ Is verified
Emotion
❑ Transcript sentiment
❑ Description sentiment

Figure 1. Operationalization of Constructs in the Theoretical Framework

2.3. Interpretable Deep Learning

Recent video analytics studies heavily utilize deep learning models, which have shown successes in object detection, video classification, and traffic monitoring, among others (Arinaldi et al. 2018, Yaseen et al. 2019). Although these deep learning models pioneer predictive analytics, their low interpretability leaves the underlying factors untapped, failing to provide actionable insights for business decision-makings.

Since the introduction of the wide and deep learning, a range of its variants have emerged. These variants fall into two categories. The first category attempts to improve the predictive power of wide and deep learning. Since the deep neural network offers the core predictive capability in wide and deep learning, studies in this category design new networks to replace the deep network, such as CNN, CRF, and attention mechanisms. The second category of variants aims to improve the interpretability of wide and deep learning. They attempt to tease out the influence of the deep component on the wide component, so that the bias of model interpretation from the wide component is mitigated.

The wide and deep learning and its variants still fall short in the following aspects. First, the wide and deep learning could only interpret the first-order relationship between the input and the output via the wide component. Although some variants attempted to model more complex relationships using other networks, the interactions among the input features are negleced. The media richness theory suggests that multiple presentation modes and their interactions are essential to model human decision-makings (Wheeler and Arunachalam 2009). Similarly, Lim and Benbasat (2002) suggest that multiple input modalities and their interactions could facilitate processing explanative information, such as information concerning relationships between or functions underlying descriptive information. Our study aims to unveil the underlying factors of misinformation transmission using multiple data modalities, including videos, audios, and texts. Therefore, modeling the interactions among these data needs to be an integral part of the method. For instance, video features and audio features could interact with each other, because good visual presentation is helpful for understanding audio messages.

Second, when training the wide and deep components jointly, the deep component affects the weights in the wide component during backpropogation. Consequently, the learned weights of the wide component (main effect) are not the total effect. Using those weights to interpret the model introduces bias caused by the deep component. Even though a few studies (e.g., Guo et al. 2020) attempted to minimize such a bias, their efforts still fail to interpret the actual and unbiased total effect.

Third, most existing methods only estimate an constant total effect for each feature, assuming the total effect is insensitive to the value changes of the feature. This assumption does not hold in real settings. For instance, when a video is only a few minutes long, increasing one minute in duration would cost the video a considerable portion of viewers. When a video is hours long, increaseing one minute in duration does not have a visible effect on its viewership. This dynamic total effect applies to many other features, though with different directions and mechanisms. To address the abovementioned limitations, we aim to devise a novel interpretable deep learning method that accounts for the interactions among multi-modal inputs, produces the unbiased total effect, and models the dynamic total effect for each feature. Table 1 summarizes the differences of our proposed method and existing methods.

3. The Proposed Approach

3.1. Wasserstein GAN with Gradient Penalty Layer

In addition to accurately predicting the transmission of health misinformation videos, we aim to estimate the total effect. GAN-PiWAD predicts the outcome variable using

$$\widehat{ADV} = \beta + \alpha_1 X_1 + \ldots + \alpha_M X_M + S(X_1, \ldots, X_M) + H(X_1, \ldots, X_M), \quad (1)$$

where $\beta + \alpha_1 X_1 + \ldots + \alpha_M X_M$ denotes the main effect, $S(X_1, \ldots, X_M)$ denotes the second-order effect, and $H(x_1, \ldots, x_n)$ denotes the nonlinear higher-order effect. The total effect of X_1 equals to the change of $\widehat{ADV}$ when X_1 increases by one unit. In order to model the dynamic total effect of each feature, we predict the total effect of each feature at every value. Let $\widehat{ADV}(X_1 = c)$ denote the expected prediction conditioned on $X_1 = c$. The dynamic total effect of X_1 under the condition of $X_1 = c$ is given by

$$\Delta\widehat{ADV}(X_1 = c) = \alpha_1 + S(X_1 = c + 1, \ldots, X_M)) - \left(S(X_1 = c, \ldots, X_M) + H(X_1 = c + 1, \ldots, X_M)\right) - (H(X_1 = c, \ldots, X_M). \quad (2)$$

The variable of interest is X_1. Therefore, the dynamic total effect of X_1 is computed as

$$\Delta\widehat{ADV}(X_1 = c) = \Delta\mathbb{E}_{X_2, \ldots, X_M}\widehat{ADV}(X_1 = c, X_2, \ldots, X_M) = \Delta\int \ldots \int_{X_2, \ldots, X_M} \widehat{ADV}(X_1 = c, X_2, \ldots, X_M) p(X_1 = c, X_2, \ldots, X_M)\, dX_2 \ldots dX_M. \quad (3)$$

However, Equation 3 is intractable because of the integral computation. In order to facilitate the computation of Equation 3, we utilize the Monte Carlo method, where the integral of a function can be approximated as the sum of function values conditioned on the samples drawn from the integrated distribution. Using this approach, Equation 3 can be transformed to:

$$\Delta\widehat{ADV}(X_1 = c) \approx \frac{1}{K}\sum_{k=1}^{K} \widehat{ADV}(x_{k,1} = c, x_{k,2}, \ldots, x_{k,M}), \quad (4)$$

where $(x_{k,1} = c, x_{k,2}, \ldots, x_{k,M})$ denotes the k-th sample drawn from the distribution $p(X_1 = c, X_2, \ldots, X_M)$. The total effect of X_1 varies as the value of X_1 changes. For visualization purposes, we also compute the average total effect of X_1 that averages over the total effects of X_1 during the value range of X_1. Assume that X_1 ranges from $c_{\min}$ to $c_{\max}$, the average total effect of X_1 is estimated as

$$\Delta\widehat{ADV}(X_1) = \frac{1}{c_{\max} - c_{\min}}\left(\Delta\widehat{ADV}(X_1 = c_{\max}) - \Delta\widehat{ADV}(X_1 = c_{\min})\right). \quad (5)$$

In order to compute the unbiased estimation of the total effect of X_1, it is necessary to learn the distribution $p(X_1 = c, X_2, \dots, X_M)$ so that samples can be drawn from it. Likewise, the dynamic total effect of $X_2, \dots, X_M$ can be calculated using the same method described in Equations 1-5.

The standard wide and deep learning cannot provide an unbiased estimation of the total effect. The standard models use the main effect from the linear part to interpret the total effect. However, the change of each feature influences the prediction from both the wide part and the deep part. Therefore, the main effect is a biased approximation of the total effect. We proposed a new method for an unbiased estimation as described above. As described above, the complexity and intractability of $p(X_1 = c, X_2, \dots, X_M)$ hinder the learning of the data distribution to draw samples. To address this limitation, we modify wide and deep learning by integrating a novel generative adversarial network (GAN) to learn the distribution $p(X_1 = c, X_2, \dots, X_M)$. GANs are a powerful class of deep generative models consisting of two networks: a generative network (generator) and a discriminative network (discriminator). These two networks form a contest where the generator produces high-quality synthetic data to fool the discriminator, and the discriminator distinguishes the generator's output from the real data. Through recurrent learning from this contest, the generator is capable of approximating the distribution of the real data. Deep learning literature suggests that the generator could learn the precise and unbiased real data distribution as long as those two networks are sufficiently powerful [2]. In order to empower those two networks and overcome the learning unstability issues of GANs, we introduce the Wasserstein GAN with gradient penalty (WGAN-GP) in this study [3]. We cohesively integrate WGAN-GP as the first layer in GAN-PiWAD. The learning loss of the discriminator (critic) in our proposed method is given by

$$L_{\mathrm{d}} = \mathbb{E}_{\tilde{\boldsymbol{x}} \sim \mathbb{P}_{\mathrm{g}}}[\mathrm{D}(\tilde{\boldsymbol{x}})] - \mathbb{E}_{\boldsymbol{x} \sim \mathbb{P}_{\mathrm{r}}}[\mathrm{D}(\boldsymbol{x})] + \lambda \mathbb{E}_{\hat{\boldsymbol{x}} \sim \mathbb{P}_{\hat{\boldsymbol{x}}}}[(\|\nabla_{\hat{\boldsymbol{x}}} \mathrm{D}(\hat{\boldsymbol{x}})\|_2 - 1)^2], \quad (6)$$

where $\mathrm{D}(\cdot)$ is a score that measures the quality of the input sample. $\mathbb{P}_{\mathrm{r}}$ is the real distribution. $\mathbb{P}_{\mathrm{g}}$ is the learned distribution by the generator. $\hat{x}$ is sampled uniformly along the straight lines between pairs of points sampled from $\mathbb{P}_{\mathrm{r}}$ and $\mathbb{P}_{\mathrm{g}}$. The distribution of $\hat{x}$ is denoted as $\mathbb{P}_{\hat{x}}$. $\mathbb{E}_{\hat{x} \sim \mathbb{P}_{\hat{x}}}[(\|\nabla_{\hat{x}} \mathrm{D}(\hat{\boldsymbol{x}})\|_2 - 1)^2]$ is the gradient penalty. λ is a positive scalar to control the degree of the penalty. The loss of the generator is:

$$L_{\mathrm{g}} = -\mathbb{E}_{x \sim \mathbb{P}_{\mathrm{g}}}[\mathrm{D}(\tilde{\boldsymbol{x}})]. \quad (7)$$

The contest between the discriminator and the generator is achieved by training Equations 6 and 7 jointly. The resulting model of this layer is a generator whose $\mathbb{P}_{\mathrm{g}}$ closely approximates the real distribution $\mathbb{P}_{\mathrm{r}}$. This generator can generate samples to compute the unbiased total effect and dynamic total effect according to Equations 1-5.

3.2. Piecewise Linear Component

The video features $\boldsymbol{X}$ are passed into the piecewise linear component, second-order component, and the higher-order component. Each feature X_j captures different aspects of a video. Within each feature, heterogeneity between different values exists as well. For instance, video creator credibility is a feature in $\boldsymbol{X}$. Videos with low creator credibility not only influence the outcome variable (misinformation transmission), but these videos may have low quality as well, which indirectly influences the outcome variable. Therefore, it is essential to consider the homogeneity among similar feature values and the heterogeneity across different feature values. Specifically, we need to differentiate the varied feature effects when the feature is at different values. Motivatied by this objective, we introduce a piecewise linear function in the linear component. For the j-th feature, let $\beta_j = \max\{x_{i,j} | i = 1, \dots, N\}$ and $\delta_j = \min\{x_{i,j} | i = 1, \dots, N\}$. We partition each feature into γ_j intervals: $[\varphi_j^0, \varphi_j^1], \dots, [\varphi_j^{\gamma_j - 1}, \varphi_j^{\gamma_j}]$, where $\varphi_j^k = \delta_j + \frac{k}{\gamma_j}(\beta_j - \delta_j)$.

3.3. Attention-based Second-Order Component

In parallel with the piecewise linear component, we devise an attention-based second-order component to model the interaction effects among the multi-modal features, as stressed by the media richness theory. The input to this component is $\boldsymbol{X}$. For each video feature $\boldsymbol{x}_i$, the interaction term of $x_{i,j}$ and $x_{i,j'}$ is denoted as $s_{i,(j,j')} = x_{i,j} \cdot x_{i,j'}$. Each interaction term has a parameter [1]. A set of M features will generate M^2 interaction terms. This will cause the learnable parameters in the second-order component to grow exponentially as the feature set increases. To prevent such an exponential growth, we propose a self-attention mechanism in the second-order component where the number of parameters is fixed. The attention-based component could scale to large number of interactions while salient interaction terms still stand out. The attention mechanism assigns a score $a_{i,(j,j')}$ to each interaction term $s_{i,(j,j')}$.

3.4. Nonlinear Higher-Order Component

The third parallel component is the nonlinear higher-order component. This component is a deep neural network that could capture higher-order effects. This network contains multiple fully-connected layers. The number of hidden layers is determined using a grid seach in the empirical analyses. The purpose of the higher-order component is to leverage the superior predictive power of deep learning to reconcile predictability and interpretability. Different from the dynamic total effect described above, the higher-order effect is a hidden component that is not interpretable, but only serves the predictive purpose. The dynamic total effect is able to delineate the magnitude of each feature's total effect at each feature value.

4. Empirical Analyses

4.1. Data Preparation

Our research testbed is collected from YouTube. A number of trusted news outlets and journal articles have identified a set of videos with health misinformation on YouTube. The topics of these videos range from COVID-19, Ebola, and cancer treatment, to vaccination. We crawled all the videos with misinformation reported by the factchecking sources. A web crawler is developed to crawl YouTube webpages of these videos for the video descriptions, video metadata, channel statistics, and detailed comments. In the end, we generated a multi-modal dataset (over 297 GB) consisting of 4,445 misinformation videos, their audios, and textual information from their webpages.

We adopt the state-of-the-art video analytics methods to extract features from each category. These features come from six data sources: video content, audio tracks, transcripts, video description, webpages, and video creators' channels. The video features are generated using the BRISQUE measurement, a widely adopted video quality measure [8]. In order to generate video features in a scalable and timely manner, we developed a python-based parallel processing method with 12 CPUs, which significantly reduced the expected computational time from 39 days to 7 days. To generate the acoustic features, we seperated the audio tracks from the videos. We utilize the Liborosa tool to compute the acoustic features [7]. In order to generate transcripts from the audio data, we developed a speech recognition model based on DeepSpeech [5]. This speech recognition model is trained on American English with synthetic noise augmentation that achieves an 7.06% word error rate on the LibriSpeech clean test corpus [9]. The trained speech recognition model was able to translate audio data into text data (transcript). The description, webpage, and channel features are extracted directly from the webpage source data. In total, we generated 854 features for each video.

4.2 Evaluation of GAN-PiWAD

We first compare with conventional machine learning methods. We repeat the training procedure of each method 10 times and report the average performance and t-test significant levels (baseline versus ours) in Table 1. We use 70% of the data for training, 20% for test, and 10% for validation. All the hyperparameters were tuned using grid search.

Our proposed GAN-PiWAD outperforms all the baseline machine learning methods in all four metrics. Compared to the best machine learning baseline model (KNN-3), GAN-PiWAD drops MSE by 5.539, MAE by 0.908, MSLE by 1.350, and MALE by 0.018. Even though KNN achieves relatively good performance among the baseline models, it is a model-free method which does not provide interpretation of the feature importance and contribution. GAN-PiWAD outperforms all the conventional machine learning models in predicting health misinformation transmission.

We, then, compare GAN-PiWAD with deep learning methods. We also test them with different hidden layers. The average performance of 10 times is reported in Table 2. GAN-PiWAD outperforms all the other deep learning methods. Compared with the best deep learning method (CNN-3), GAN-PiWAD reduces MAE by 0.90, MSLE by 0.046, and MALE by 0.093. Our proposed GAN-PiWAD remains the best in prediction accuracy when the hidden layers of the baseline models change.

Since GAN-PiWAD is an interpretable method, we select the state-of-the-art interpretable methods for comparison. The average performance of 10 times is reported in Table 3. Compared with the best interpretable model (W&D), our GAN-PiWAD reduces the MSE by 23.347, MAE by 0.436, MSLE by 0.09, and MALE by 0.08. Compared with other interpretable methods, GAN-PiWAD consistantly obtains the best performance. The performance improvement is attributed by the architecture of GAN-PiWAD.

Table 1. Comparison of GAN-PiWAD with Conventional Machine Learning

Method	MSE	MAE	MSLE	MALE
GAN-PiWAD (Ours)	157.522	5.579	0.977	0.739
Linear regression	881.027***	12.812***	3.184***	0.891***
KNN	180.479	6.309*	2.264***	0.758
DT	284.387***	8.433***	3.362***	0.901***
SVR	185.644	9.483***	4.924***	1.267***
Gaussian Process	1291.331***	23.791***	8.508***	1.579***

Table 2. Comparison of GAN-PiWAD with Deep Learning Methods

Method-Layer	MSE	MAE	MSLE	MALE
GAN-PiWAD (Ours)	157.522	5.579	0.977	0.739
MLP	181.192	6.128	0.932	0.854*
CNN	169.584	6.633**	1.065	0.849**
LSTM	341.301***	9.715***	1.718***	0.973***
BLSTM	367.261**	9.728***	1.661***	0.973***

Table 3. Comparison of GAN-PiWAD with Interpretable Deep Learning

Method	MSE	MAE	MSLE	MALE
GAN-PiWAD (Ours)	157.522	5.579	0.977	0.739
W&D [1]	180.869	6.015	1.067	0.819*
W&D-CNN [6]	186.773	6.304*	1.866*	1.039**
W&D-LSTM [10]	183.719	6.141	1.598	0.943
W&D-BLSTM [11]	206.321	6.648*	2.454***	1.210***
Piecewise W&D-10 [4]	227.633*	7.126***	3.116***	1.420***
Piecewise W&D-20 [4]	206.792*	6.805**	3.016***	1.395***

*: p < 0.05; **: p < 0.01; ***: p < 0.001

5. Interpretation of GAN-PiWAD

Following the standard approach from previous studies, we plot the main effect of the baseline methods to interpret their predictions. Such an interpretation from the main effect is a biased approximation because of the influence from the high-order component. For visualization simplicity, we average over all the video features into one feature, as they all represent the video quality with the same scale. We also average over all the acoustic features into one feature, because they measure the audio quality with the same scale. In order to compare all the features in the same scale, we normalized the effect values in Figure 2.

As shown in Figure 2, the rewards features have the most salient influence on the prediction, especially the description features. The description is a brief paragraph presented below a video describing an overview of the video content. This is usually the first content a viewer would read about the video, which directly influences whether the viewer will actually watch the video. The description features are the number of medical terms in the description, informativeness, readability, and complexity of the description, among others. The results show that one unit of increase in description readability results in 88.57 units of increase in average daily views. The number of medical terms has the most influence on the prediction. One unit of increase in the number of medical terms in the description will raise the average daily views by 138.69 units. These features measure how well the description can be perceived and how much medical information it contains. A easy-to-read and medically informative description leads to more transmission of health misinformation as the viewers attempt to seek medical information from the videos. Conveying the medical information that the viewers wanted to the largest extent could entertain the viewers and retain them to watch the rest of the video. If the medical information is easy to comprehend, the viewers have a better understanding of the video topic, which motivates them to watch the details from the video.

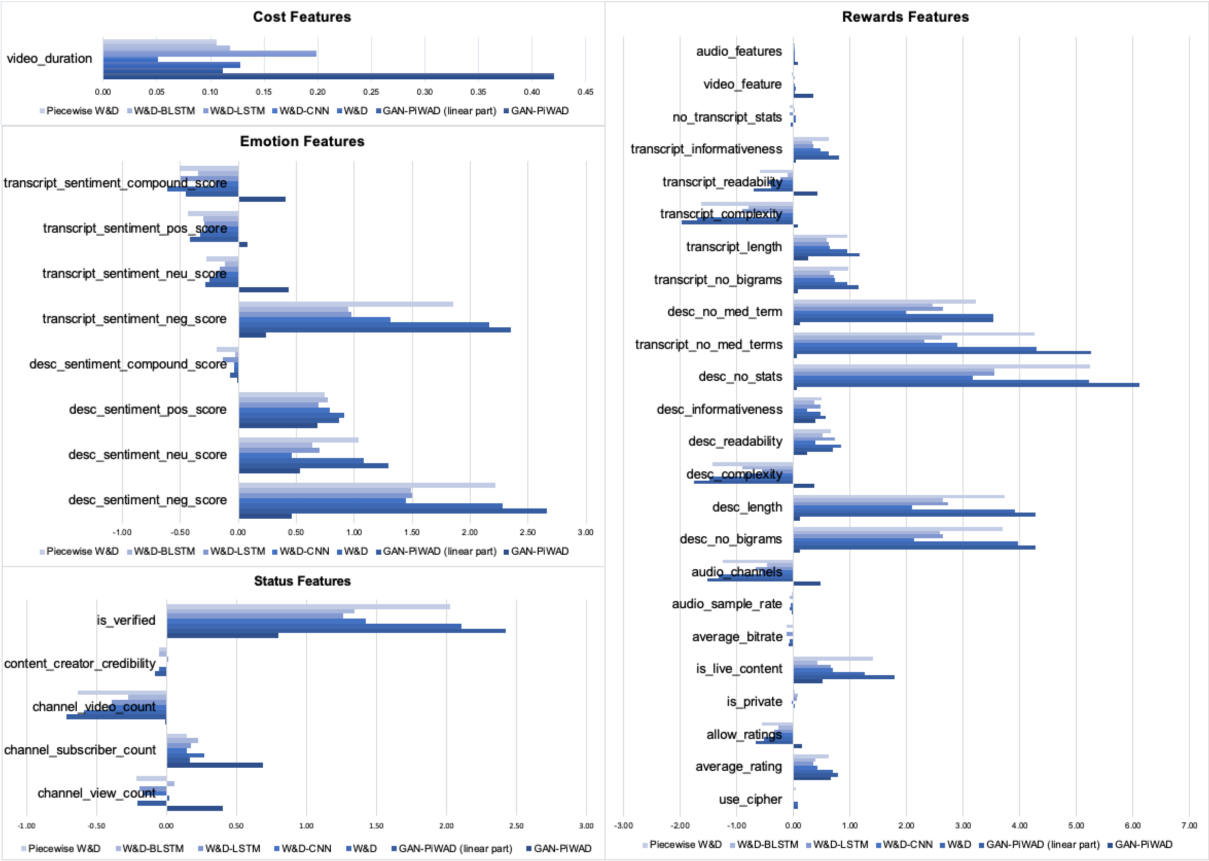

Figure 2. The Effect of Each Feature on the Prediction (Normalized)

The transcript and description sentiments also significantly affect the transmission of misinformation. Notably, the negative sentiment expressed in the description and transcript has the most significant influence. A unit of increase in the description negative sentiment score (indicating stronger negative sentiment) leads to an increase of 0.36 units in average daily views. A unit of increase in the transcript negative sentiment score increases the average daily views by 0.67 units. Negative sentiments in video content increase the transmission. In the context of health misinformation, many videos contain very emotional messages from narrators who report their experience of using medical products. For instance, some described the autism diagnosis of their children after receiving vaccination. Others reported the risk of using WiFi as they believe it is linked to coronavirus infection. The myth that eating certain food would cause cancer is also commonly shared in the misinformation videos. These negative emotions and personal narratives escalate the viewers' opposition to vaccines and food types without a factual base.

The status features of the channel have a critical influence on the transmission as well. In particular, if a channel is verified, the transmisibility of misinformation is higher (increase the average daily views by 0.8 units). YouTube collects information from verified channels, such as phone numbers. Verified channels signal authenticity and credability to viewers. Therefore, the viewers are more likely to watch the videos posted by these channels, regardless of whether they are misinformation or not.

The interpretation of the prediction sheds light on the management of infodemics for video platforms. These platforms could utilize our method to monitor the description features. Medical-related videos whose description is well perceived should be under scrunity. When a video shows overwhelmingly negative content, it needs to be closely monitored as well to prevent misinformation spread widely. Special consideration should be given to verified video channels, because their videos have higher likelihood to transmit easily than other channels.

6. Discussion

6.1. Methodological Implications

We devise a novel GAN-PiWAD method to predict health misinformation transmission and interpret such a prediction. GAN-PiWAD innovatively incorporates the interaction effect into the model, unlocks the possibility to estimate the unbiased total effect, and captures the dynamic total effect for each feature. Our method augments an attention-based interaction branch to the wide and deep framework. The new framework learns three components jointly (piecewise linear, attention-based interaction, and higher-order components). In order to tease out the influence from the high-order component, we introduce a Wasserstein generative adversarial network with gradient penalty (WGAN-GP) within the wide and deep model. Our proposed method outperforms strong baseline models. GAN-PiWAD is not restricted to the misinformation transmission prediction context. It is an generalizable interpretable method to understand the underlying factors of human behavior, including healthcare, cybersecurity, and technology acceptatnce, among others.

6.2. Practical Implications

This study offers many practical implications for the stakeholders. For the social media platforms, our method is an implementable analytics tool that can predict widely transmissible misinformation. Our method also offers the interpretation of the prediction. Critical features are identified to understand the transmission. To prevent this misinformation from spreading, the social media platforms could utilize our research findings to design intervention measures. For instance, negative videos from verified channels with easy-to-read descriptions need to be specially monitored for misinformation. For the health sectors, our method and research findings open a door to manage infodemics. Containing infodemics could significantly alleviate the burden of the health sectors to control pandemics. For the policymakers, we offer an automated tool to identify misinformation. The policymakers could utilize our method to trace major misinformation sources and hold them accountable.

6.3. Limitations and Future Directions

First, besides understanding misinformation transmission, our method could also understand other human behaviors in healthcare, cybersecurity, and technology acceptance. Future work could test the efficacy of our method in other research contexts. Second, we focused our empirical analyses on YouTube misinformation vidoes. Other social media platforms, such as Facebook and Twitter, are also popular outlets for misinformation. To confirm our research findings, more ground truth data could be collected from Facebook and Twitter to perform the empirical analyses. Fourth, since there is no standardized quantitative measurement of interpretability, we visualized the model interpretation of our method and other interpretable methods. Future studies could design theoretical and empirical frameworks to quantify the interpretability of a method, so that the comparison between interpretable methods is clearer.

7. References

[1] Heng Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, Rohan Anil, Zakaria Haque, Lichan Hong, Vihan Jain, Xiaobing Liu, and Hemal Shah. 2016. Wide & deep learning for recommender systems. In *ACM International Conference Proceeding Series*, 7–10.

[2] Ian Goodfellow, Joshua Bengio, and Aaron Courville. 2016. *Deep Learning*.

[3] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. 2017. Improved Training of Wasserstein GANs. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, 5767–5777.

[4] Mengzhuo Guo, Qingpeng Zhang, Xiuwu Liao, and Daniel Dajun Zeng. 2020. An interpretable neural network model through piecewise linear approximation. (January 2020).

[5] Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, and Andrew Y. Ng. 2014. Deep Speech: Scaling up end-to-end speech recognition. *arXiv* (December 2014).

[6] Jun Woo Lee and Yeo Yoon Chan. 2019. Fine-Grained Plant Identification using wide and deep learning model. In *2019 International Conference on Platform Technology and Service, PlatCon 2019 - Proceedings*.

[7] Brian Mcfee, Colin Raffel, Dawen Liang, Daniel P W Ellis, Matt Mcvicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and Music Signal Analysis in Python. In *PROC. OF THE 14th PYTHON IN SCIENCE CONF.*

[8] Anish Mittal, Anush Krishna Moorthy, and

Alan Conrad Bovik. 2012. No-reference image quality assessment in the spatial domain. *IEEE Trans. Image Process.* 21, 12 (2012), 4695–4708. DOI:https://doi.org/10.1109/TIP.2012.2214050

[9] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An ASR corpus based on public domain audio books. In *IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 5206–5210. DOI:https://doi.org/10.1109/ICASSP.2015.7178964

[10] Nail Tosun, Egemen Sert, Enes Ayaz, Ekin Yilmaz, and Murat Gol. 2020. Solar Power Generation Analysis and Forecasting Real-World Data Using LSTM and Autoregressive CNN. 1–6. DOI:https://doi.org/10.1109/sest48500.2020.9203124

[11] Hongfan Ye, Buqing Cao, Zhenlian Peng, Ting Chen, Yiping Wen, and Jianxun Liu. 2019. Web Services Classification Based on Wide & Bi-LSTM Model. *IEEE Access* 7, (2019), 43697–43706. DOI:https://doi.org/10.1109/ACCESS.2019.2907546